Assessment of Data After Dimensionality Reduction

ABSTRACT

PURPOSE: Implement Principal Component Analysis (PCA) to compute the Singular Value Decomposition (SVD) to reduce the dimensions of the data. Specifically, we will be transforming 13 Dimensions into 2 Dimensions, while using the PCA to determine the most effective reduction. We will identify the optimal pair values using the Davies-Bouldin index provided in the assignment.

METHODS: In this study, we used MATLAB to process the "wine.csv" dataset. The optimal pair of components for the most effective reduction was identified based on the lowest DB index score. The produced lowest DB value and the column combination were presented along with a visualization using gscatter. We then performed the reduction using SVD on the zero-mean of the data(Part B) and then repeated the process on standardized zero-mean data. (Part C) In parts B and C the numerical score of the DB index and a visualization using gscatter were also produced

RESULTS: Running the program produced 3 figures that show the clustering of the 3 cultivar groups. It allows us to compare the effectiveness of the three different clustering methods used. We can use visual inspection and conclude using the standardized PCA was the most effective as the clusters have the most clear separation of the 3 figures.

CONCLUSIONS: The three methods that were implemented and the computed DB Score were all shown to be within expectations. It should be noted that the dimensionality reduction from 13D to 2D performed the worst both in visual inspection of the figure and the highest DB score. The values found using the DB score coincide with the visual inspection with standardized having the lowest followed by “best” pair with the reduction using PCA being the worst.

INTRODUCTION

SCIENTIFIC PURPOSE/OBJECTIVE: The scientific objective is to implement and evaluate the effectiveness of dimensionality reduction using PCA with zero mean data and the DB score.

BACKGROUND: The use of dimensionality reduction is vital to data analysis of any large or complex dataset. As datasets grow larger and more complex they become increasingly harder to interpret, for both humans and AI as training on overly large and complex data causes an increase in computational cost and overfitting of the model (Shrivastava) This is where dimensionality reduction techniques such as the one we’re working on, Principle Component Analysis, can help. Reducing the number of variables in the dataset these methods attempt to retain the most important – principle – information while mitigating the issues caused by overly large datasets.

PCA, the purpose of the method we are focused on for this assignment is to “reduce the number of variables of a data set while preserving as much information as possible.” (Jaadi)

The implementation we’re using involves computing the singular value decomposition (SVD) of a data matrix where the mean of each column is zero, the top eigenvectors will then be selected to represent the whole of the data as principle components.

The Davies-Bouldin index used in this assignment for evaluating the effectiveness of the clustering methods is provided by the instructor for this assignment. Briefly explained this index is a validation metric used to evaluate clustering models. It is calculated using the average similarity measure of each cluster with the cluster most similar to it. It used the average distance between points in a cluster and the centroid and compared against the distance to other clusters. In general, the lower score given through this index means a better clustering method.(GfG)

We used our three clustering methods and the provided DB index score implementation to explore which method is best for clustering the given dataset.

SCIENTIFIC QUESTION:
Given the wine.csv dataset, is the dimensionality reduction using PCA effective at clustering the data with DB score serving as a performance metric?

We will test this by implementing the three different variations of dimensionality reduction using PCA in MATLAB, with the provided DB score implementation we can use the generated figures and score to evaluate the effectiveness of clustering given each variation.

METHODS

The first step is the evaluation and preprocessing of the data, upon inspecting the data visually by opening it it can be seen that the 13 dimensions/independent variables are in the first column and are shown by each row. During preprocessing the data was transposed before separating into Xmat and lvec, this made the 13 dimensions of each column. With this, the data was ready for analysis.

Problem A required us to identify the columns that would provide the most effective clustering of the data by finding the pair of columns that had the lowest DB score. To obtain this a nested for loop was created to find combinations of every column in which the resulting DB score calculation was stored in a matrix of equivalent size. We were then able to take the stored scores and take the smallest value along with the columns that were used to obtain that calculation.To visualize the clustering we’ve obtained we use gscatter to create a scatter plot.

For problem B, PCA was applied to the data to reduce the 13 dimensions/variables to 2 dimensions. The first step in the process is to zero-mean the data. Next SVD was performed on the resulting zero mean data to obtain the right singular vectors. This was then used to project the data onto the first two principle components resulting in the desired 2D representation of data.

Once again the DB score was obtained as well as the use of gscatter to create a scatter plot of the clusterings.

For problem C, the data (Xmat) is first standardized using the zscore function in MATLAB before applying the same steps as in problem B. The standardization ensures the scale of the variables is the same and removes any differences in units, which is important as the weight of certain variables could be too high simply because the numbers are larger due to the unit it was measured in. After this the DB score was calculated and the clusters were plotted using gscatter.

RESULTS

Table 1: Results of Davies-Bouldin Index. Test column refers to the different methods used. DB Index is the score of the method. Variables column is the columns found for data columns method.

Test	DB Index	Variables
Data Columns	0.78748	[7 1]
Raw PCA Scores	1.5148	
Standardized PCA	0.6392	

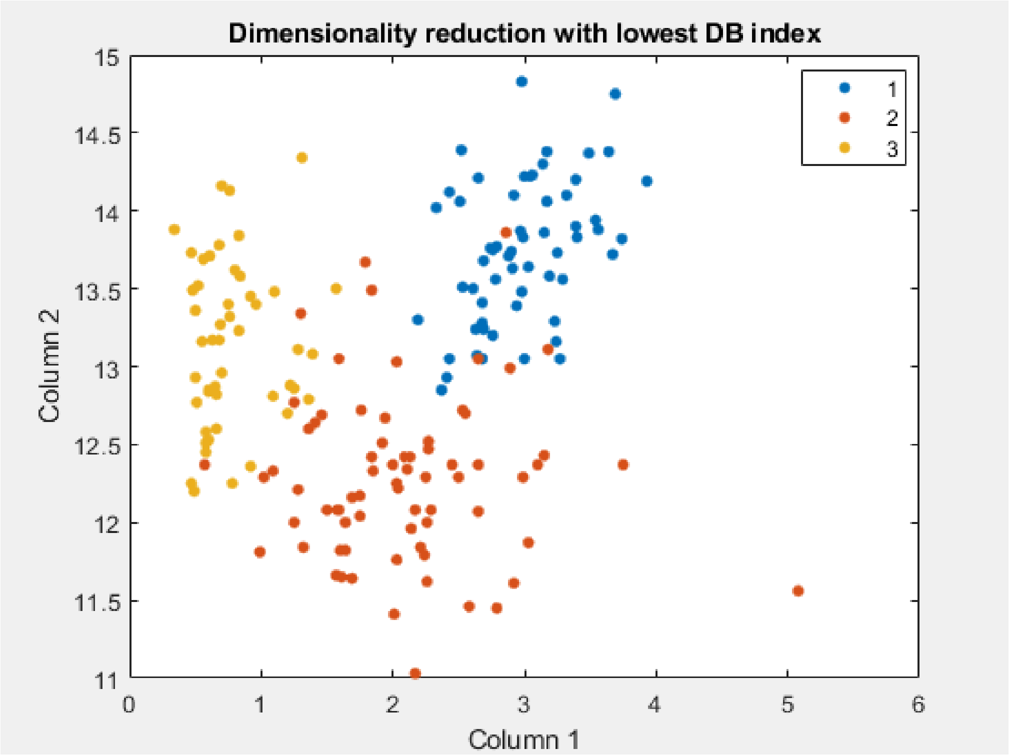

Figure 1: Scatter plot using pair of columns that provide the “best” dimensionality reduction according to DB Index

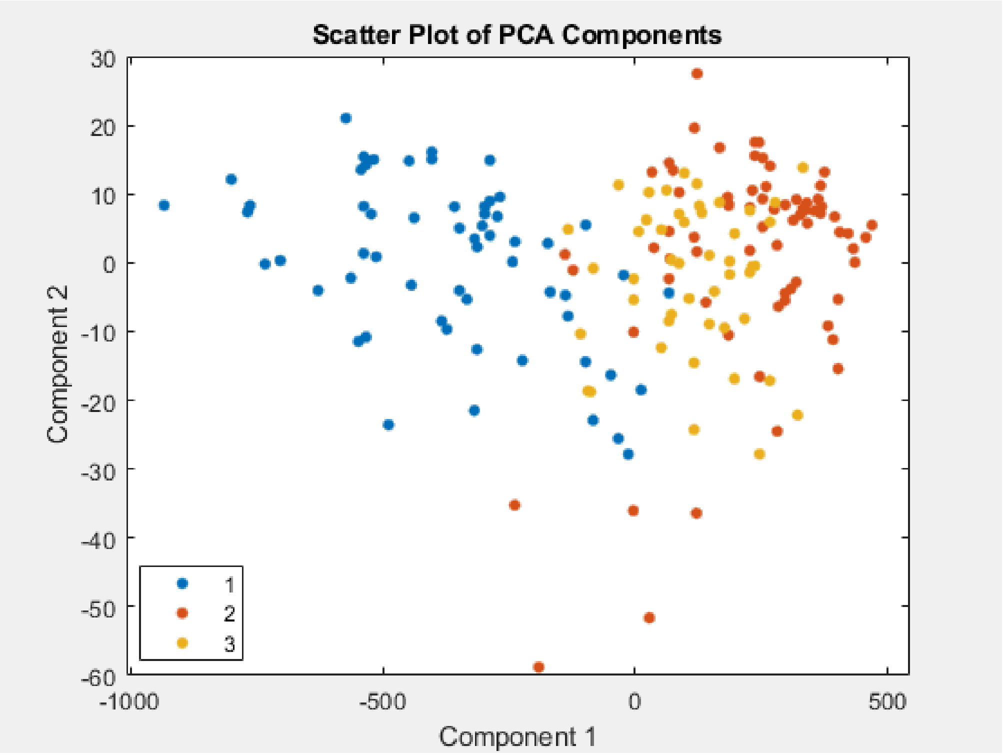

Figure 2: Scatter plot after reducing the 13D data to 2D data using PCA

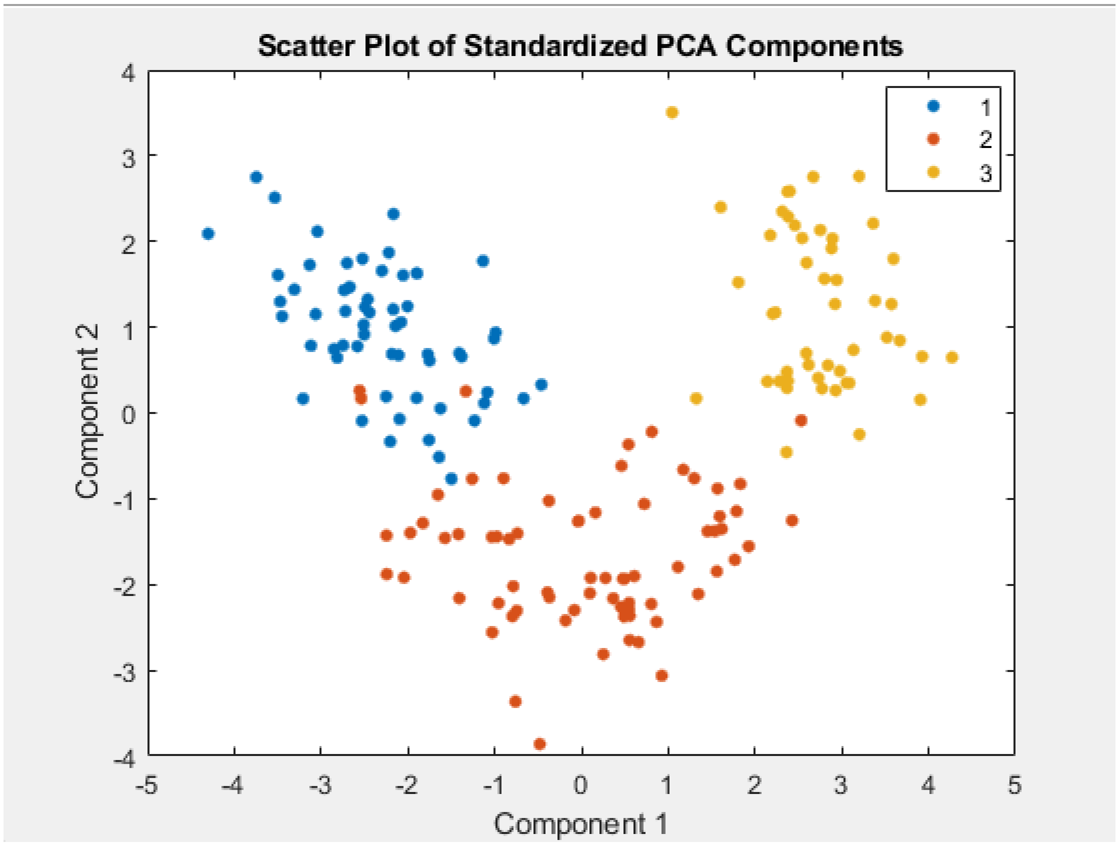

Figure 3: Scatter plot using standardized data reducing using PCA

DISCUSSION

The result presented in the PCA SVD analysis provides insights into the effectiveness of using dimensionality reduction on a given dataset. The Davies-Bouldin Index score served as an easy and objective way of quantifying the quality of a given clustering method. As basing the quality of clustering based on only visual inspection is prone to error. With DB Score we can safely assume the lower the score, the better.

In all problems A-C we observed that visual inspection of the clustering matched the DB scores that were generated, with this we can explore each problem/method of clustering.

In problem A, the DB score was utilized to identify the two columns that provided the most effective data clustering. It found columns 7 and 1 to have a DB score of 0.78748 which indicates that those two columns were the best two columns suited to represent the dataset. This information provides some insight into the nature of the data, perhaps these two columns/variables serve as aggregates of the other data or the nature of it takes into account the other elements. They could also have some information that can help distinguish patterns within the data. We should however note that unlike PCA there does not seem to be any attempt at retaining or minimizing information loss with this method and important information could be lost while utilizing this method.

In problem B, the PCA method was utilized while performing dimensionality reduction. It provided a DB score of 1.5148 which indicates a rather poor clustering quality compared to Problem A's method. This could be due to the lack of standardization which we will do in Problem C, without standardization we run into an issue of the scaling between variables. With different units of measurements some variables could have numbers in the millions while others in the decimals, this would cause the weighting of the larger numbers to obscure the effect of all other variables resulting in the poor clustering we see.

In problem C, the data was standardized before performing the PCA method for dimensionality reduction. It provided a DB score of 0.6392 which is the lowest we've seen as well as significantly lower than problem B where the only difference is the standardization of the data. This indicates standardization has had a significant effect on the quality of clustering for this dataset. This contrast in clustering quality highlights the importance of data preprocessing before attempting to apply the PCA dimensionality reduction. This drastic change also indicates that the previously mentioned difference in scale is most likely present within this dataset.

Overall the results found within the three methods explored in this assignment show the significance of proper data preprocessing and employing the proper technique to improve the quality of data analysis.

REFERENCES

GfG. "Davies-Bouldin Index." *GeeksforGeeks*, 5 Nov. 2023, www.geeksforgeeks.org/davies-bouldin-index/. Accessed 27 Feb. 2024.

Jaadi, Zakaria. "A Step-by-Step Explanation of Principal Component Analysis (PCA)." *Built In*, 2023, builtin.com/data-science/step-step-explanation-principal-component-analysis. Accessed 27 Feb. 2024.

Shrivastava, Dr. Virendra Kumar. "Why Dimensionality Reduction Is Crucial in Machine Learning Models?" *LinkedIn*, 11 May 2022, www.linkedin.com/pulse/why-dimensionality-reduction-crucial-machine-learning-shrivastava/. Accessed 27 Feb. 2024.